SOME THOUGHTS ON THE USE AND MISUSE OF STATISTICAL INFERENCE

Ralph E. Strauch

January 1969

P-3992

SOME THOUGHTS ON THE USE AND MISUSE OF STATISTICAL INFERENCE

Ralph E. Strauch
The RAND Corporation, Santa Monica, California

INTRODUCTION

There is strong pressure within government and within the society as a whole for quantitative analysis of an ever widening class of problems, in order to produce "objective" results in a "scientific" manner. Such analyses, it is hoped, will minimize the extent to which the conclusions drawn depend on the "subjective" judgment of the analyst drawing them, and maximize the extent to which they reflect "objective" reality. This pressure has produced an infatuation with computational techniques coupled with a neglect of the conceptual principles underlying the development of those techniques which sometimes results in their application in situations in which the interpretation given to the results, if not the techniques themselves, are inappropriate.[1] Notable examples of this have occured in studies arising from the Vietnam war and in the study and interpretation of various forms of domestic data, such as criminal statistics.

One form of analysis which seems to me to be abused with great frequency is statistical inference. Whenever data is available in a form similar to that which would be obtained from a process for which a standard statistical model and well defined analytical techniques exist, and the questions of interes about the real world process producing the data are similar to questions about the standard model which can be answered with accepted techniques, it

is tempting to apply these techniques to the data and to interpret the results as though the model were an adequate representation of the real world process. This is often done even though the analyst involved is fully aware that this standard model is not a reasonable one for the process being studied and that the interpretations being made have objective meaning only in the context of that model.

Consider, for example, the problem of comparing the fraction of two populations having a particular characteristic based on data obtained from a fortuitous sample of those populations. Standard techniques are available for testing the hypothesis that the fraction of a given type of individual is the same in two populations, given a random sample from each of them, and it seems natural to use these techniques. But if we then find that the difference is significant, at say, the .05 level, what have we learned? The statement that the results are significant at the .05 level means that, if the samples were obtained by random sampling from two populations each with the same fraction of individuals of the type in question, then the differences between the samples as great as or greater than that actually observed would be expected to occur 5% of the time. It is only a statement about what we would expect given random sampling, and says nothing directly about what we should expect if the data were produced in a different manner. If the data available cannot

reasonably be considered a random sample from anything, statements about the statistical significance of differences between groups occurring in that data provide, by themselves, little objective evidence about that difference. In such circumstances, the best interpretation which I can find for the statement that the data are "statistically significant" is that the author has shown that it is unlikely that the data were produced by a particular process (in this case, random sampling from identical populations), which is often one which no reasonable man would believe is the process producing them, therefore he would like the reader to accept his explanation of the data rather than other explanations which he has not seriously considered.

Inferences drawn in this manner cannot be considered "objective," since their validity in the real world depends heavily on the relationship between the real world process being investigated and the model from which the inferences are derived. Judgments are required about the nature of this relationship, and about the way in which results should be interpreted given the discrepancies between the model and the process being studied. These judgments are seldom made explicitly, and in fact, are sometimes not even given serious consideration.

Statistical inference, or any other form of quantitative analysis, provides truly objective results only in very special circumstances,

for example, when the analyst is able to mold the process being investigated to conform to the model used, rather than the reverse. This is, after all, what occurs in a sample survey from a well defined finite population using probability sampling methods, or in other experiments in which randomization is introduced by the experimenter. In the analysis of the data produced as a fortuitous byproduct of some other activity, however, statistical or other quantitative techniques can be at best an aid to judgment, and if they are to serve this function well, this judgment should be careful and explicit.

This is not to imply, of course, that conclusions reached through the application of statistical techniques to data when the assumptions on which those techniques are based are not satisfied are necessarily wrong. That certaintly is not the case, anymore than conclusions reached through the use of a crystal ball, astrology, or flipping a coin are necessarily wrong. If I flip a coin to determine the answer to all my yes–no questions, I will, after all, be right about half the time. The issue, then, is not whether a particular conclusion is correct or incorrect, but rather it is the basis for confidence in that conclusion. It is, it seems to me, reasonable to base confidence in a conclusion on a particular theory or model only to the extent to which we are willing to accept the axioms or other assumptions (explicit or implicit) of the theory or of the model, and the logical consequences which follow from them. My concern over

some of the current uses of statistical inference (and other forms of decision analysis as well) stems from the fact that conclusions are given theoretical justification when in fact the assumptions required for this justification are obviously violated, and the degree of violation and its effects are either ignored or glossed over.

AGREEMENT BETWEEN THE MODEL AND THE PROBLEM

In problems of statistical inference, lack of agreement between the model and the problem being modeled can occur on two different levels, violation of particular detailed assumptions inherent in the specific model being used, and violation of the basic principles and concepts relating to decision in the face of uncertainty which are implicit in the theory of statistical inference and in decision theory in general. The first type of violation is usually easy to identify, and occurs almost invariably to some degree. The model is only an approximation of reality, and there may even be approximations within the model itself (e.g., approximation of a binomial distribution with a normal). In most instances, the effects of such approximations may not be particularly harmful. On the contrary, the use of approximations is quite helpful, by eliminating or reducing tedious computations which would otherwise be required or by bringing out the effects of important elements or factors in the problem more clearly than would be possible with a more complex model. As either the model or the problem being modeled becomes increasingly complex, however, it becomes increasingly harder (and more important) to keep track of the effects of the approximation involved, and to take account of them in interpreting the results. The second type, violation of principles implicit in the theory, is more fundamental, and when it occurs it may cast considerable doubt on usefulness of

the analysis involved. There is not always a clear dividing line between the two, but instead there may be a gradually increasing scale of disagreement between the specific assumptions inherent in the model and reasonable assumptions about the real world such that as we move further up the scale it becomes increasingly more difficult to reconcile the application of the model to the real problem with the basic principles of statistical inference.

Statistical inference, in principle, never involves direct inference from the data observed to the process causing the data (e.g., from the sample to the population in the case of sampling). It consists, instead, of comparing the observed data with that expected from various members of a collection of predictive models which are assumed to be adequate models of possible alternative versions of the process being observed. The manner in which the results of this comparison are used in the subsequent inference depends on the nature of the inference being made and on other criteria chosen by the analyst. For purposes of this discussion, however, the important point is that observed data can be interpreted relative to the alternative versions of the process represented in the predictive models considered, and only relative to those. If the collection of predictive models considered does not provide a reasonable (and this can be decided only through the careful use of subjective judgment) representation of the real world alternatives

of interest, then the theoretical considerations which provide justification for the inference produced by the model (of which this collection of predictive models is but a part) provide no justification for the interpretation of that inference in the real world problem.[2]

THE USE OF PREDICTIVE MODELS

To see the way in which predictive models are used in the development of statistical techniques, let us turn to a standard statistical example, that of sampling from an urn filled with colored balls.[3] Assume we have an urn containing some fixed number of balls, say 100, and that an unknown number (possibly zero) of these are red and the rest are black. We draw a random sample of balls from the urn, and wish to use this sample as a basis for inference about the composition of the urn. The observed data in this case is a description of the sample, and the process producing the data consists of sampling, according to a fixed sampling procedure, from an urn of fixed (but unknown) composition. The class of predictive models which is considered, therefore, should include a predictive model which describes the way in which each possible urn composition and sampling procedure will produce samples.

If the sample is obtained by random sampling without replacement (i.e., when a ball is drawn from the urn it is not replaced and so has no chance of being drawn again), and we wish to make no a priori assumptions about the composition of the urn, we should then consider 101 different predictive models, corresponding to the 101 different possible urn compositions (of r red balls and 100–r black balls, $0 \leq r \leq 100$), and our method of inference should take into account the way in which samples are produced by each of

these predictive models.[4] Standard statistical techniques applicable to the problem do, in fact, do this. Fortunately, this is easier than it may appear at first glance, because each of these predictive models results in the production of samples according to a distribution of the same form (hypergeometric), and we can handle them all in the form of a single general model with a varying parameter within the model.[5] We should, however, be aware that in doing this we are, conceptually at least, dealing not with a single model or single process, but with 101 different predictive models of 101 different versions of a process only one of which is actually occurring when we draw the balls from the urn.

The assumption of random sampling, then, and the use of techniques based on that assumption, results in inferences based on the comparison of the observed data with the expected outcomes produced by predictive models of the results of random sampling from urns of all possible compositions. If the true process producing the data is not represented, i.e., if the balls are obtained by some means other than random sampling, and if the relationship between the actual urn composition and the sample produced by this process is significantly different than the relationship for random sampling, the use of the random sampling model may lead us to totally inappropriate conclusions, and will certainly lead to inferences whose performance characteristics (probability of error, etc.) are quite different from what we expect.

But how restrictive is the assumption of random sampling? To answer this question we must contrast normal usage of the term random with the much narrower and more restrictive technical definition of "random sampling" in the statistical sense. Webster defines the adjective "random" as meaning "lacking or seeming to lack a regular plan, purpose, or pattern," and also gives the additional definition "having the same probability of occurring as every other member of a set." The former definition, while consistent with ordinary usage of the term random, is in contradistinction to the statistical definition of the term. Random sampling, in the statistical sense, imposes the positive requirement on the sampling procedure that every possible sample be equally likely, and this is far different from the essentially negative requirement that there be no obvious pattern or plan. The latter definition, if interpreted properly, encompasses the statistical meaning, but if interpreted improperly, may also imply something far different. The set whose members must have the same probability of occurring is not the set of balls in the urn, but rather it is the set of all possible samples (of the size we are considering) of such balls. Thus, for example, if we number the balls from 1 to 100 then choose a digit from 0 to 9 at random and take, as a sample of size 10, all balls whose number ends in the digit chosen, we do not obtain a random sample. While it is true that each ball has equal probability of being in the sample

(.1), it is not true that each sample is equally likely, since we will never obtain a sample containing, say, both ball 16 and ball 48.

The assumption of random sampling is, therefore, quite a restrictive one. In applications to real data, it imposes a positive requirement on the analyst to insure that his sampling procedure does indeed provide a random sample or a reasonable approximation to one, and not simply a negative requirement to insure that there are no obvious or conscious sources of bias. It is not sufficient that each ball in the urn have equal probability of being included in the sample, but it must also be true that each subset of balls have equal probability of being included. That the assumptions of the models employed be satisfied is not simply a matter of mathematical nicety, but is a fundamental to the validity of the inference, since the inference is only a statement about what is reasonable <u>if</u> the process being observed is one of those represented by the collection of predictive models employed. Moreover, this is true whether or not the analyst makes explicit use of the predictive models, since even if he uses only techniques explicitly, he is implicitly using the collection of predictive models for which those techniques were developed.

The problem would be less serious if the models were self verifying, in the sense that if the model did not fit the process

producing the data, this would be evident from the data and would prevent incorrect inferences from being drawn. Unfortunately, this is seldom the case. The collection of predictive models contained in the statistical models of many common statistical problems is large enough to explain almost any observed data to which the model is applied. In the case of balls drawn from an urn, for example, the collection of predictive models representing random sampling from all possible urn compositions is sufficient to explain any sample of balls, so that there is no way to determine whether or not a sample of balls drawn from an urn of unknown composition was drawn at random on the basis of the sample alone. (We could, on the other hand, make inferences about the randomness of the sampling procedure given a sample of balls from an urn of known composition. To do this we would use predictive models representing alternative versions of the process of sampling from the urn of fixed composition according to the set of alternative sampling methods we wishes to consider. In fact, given any two of the three elements of the problem, the urn composition, sampling procedure, and resulting sample, it is possible to make meaningful statements or to draw inferences about the third. If we have known only one of the three, however, there is little we can say about the other two.)

It is, of course, possible to draw valid inferences from nonrandom samples. In fact, the practical difficulties involved in

random sampling from a very large population are such that random sampling is seldom used in such cases. When a different sampling procedure is used, however, the predictive models on which the inferences is to be based should reflect the procedure actually used and not simple one which tends to be mathematically convenient.

The basic principle underlying all statistical inference is that we attempt tc distinguish the process actually being observed from alternative possible versions of that process on the basis of expected differences in the outcomes produced by these versions. The use of predictive models which do not describe the behavior of the alternatives among which we wish to distinguish, or of techniques based on such models, is a clear violation of this principle.

REFERENCES AND NOTES

1. I am not condeming the use of statistical techniques to "snoop" through large amounts of data to look for possible interesting relationships worthy of further study. Such "snooping," however, is not statistical inference, and relationships thus found should not be interpreted as though it were.
2. I am not concerned in this discussion with other forms of justification often used in practice, such as "Everyone does it this way and it seems to work," "We have all this data and have to do something with it," or "This is what they said to do in Stat 309B," as these make no use of, and have no bearing on, any theoretical considerations related to the validity of the inferences drawn.
3. This discussion illustrates an important, and sometimes overlooked, distinction between mathematicians and most other scientists in the way in which they relate mathematical models to the real world. Most scientists are concerned primarily with the real world, and use models to help them understand it. Mathematicians, on the other hand, are often primarily concerned with the models, and use the real world to gain a better conceptual understanding of the model. In this discussion I am not really interested in urns filled with red and black balls, but in a class of models which seem to describe some of the important aspects of sampling from a

finite population. The real world picture which I draw of these models (the urn) is simply to aid in understanding them.

4. If we do make a priori assumptions about the urn composition (such as "at least half of the balls are red"), the composition of collection of predictive models required will be reduced accordingly.

5. For a discussion of the mathematical details, the reader should refer to any elementary statistics text, such as Hodges and Lehmann, Basic Concepts of Proability and Statistics, Holden-Day, 1964.